DOCUMENT RESUME

ED 441 813 TM 030 847

AUTHOR Kiplinger, Vonda L.; Haug, Carolyn A.; Abedi, Jamal
TITLE Measuring Math--Not Reading--on a Math Assessment: A Language Accommodations Study of English Language Learners and Other Special Populations.
PUB DATE 2000-04-00
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Elementary School Students; Intermediate Grades; *Limited English Speaking; Mathematics Achievement; *Mathematics Tests; *Reading Ability; Spanish Speaking; State Programs
IDENTIFIERS *Colorado; *Testing Accommodations (Disabilities)

ABSTRACT

In 1998 it became apparent that support for state assessments in Spanish at the elementary level was waning in Colorado, and it was likely that only an English-language version would be available for the grade 5 mathematics assessment in 1999. A study was conducted to provide empirical results to inform decisions regarding language accommodations for students who are English language learners from all language backgrounds and students with special needs. Three test forms of the mathematics test were administered to 1,198 fourth graders in 1999: (1) original English; (2) simplified English; and (3) original English with glossary. The initial conclusion that can be drawn from these data is that performance on mathematics assessments with high proportions of word problems is directly related to proficiency in reading English. Examination of mathematics performance within Language Assessment Scale categories (measure of English proficiency) shows that simplification of linguistic structures and the addition of a glossary for nonmathematics vocabulary to a mathematics assessment results in better performance by English language learners and other students who are not good readers. Data suggest that linguistic simplification or clarification of the vocabulary of mathematics word problems can benefit virtually all students. (Contains 8 tables and 30 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 813

Measuring Math -- Not Reading -- on a Math Assessment: A Language Accommodations Study of English Language Learners and Other Special Populations

by

Vonda L. Kiplinger
Carolyn A. Haug
Jamal Abedi

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

V. Kiplinger

TM030847

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 24-28, 2000.

Introduction and Purpose of the Study

In 1993, the Colorado General Assembly passed the Education Reform Act which established a system of standards-based education reform. The twin cornerstones of this reform effort were the development of content standards and assessments that "reflect the highest possible expectations" – the goal of which was to "enable today's students of all cultural backgrounds to compete in a world economy in the twenty-first century." (*Colorado School Laws 1998, Part 4, Educational Reform, 22-7-401*, p. 106). This *Act* established the Colorado Student Assessment Program (CSAP) for the administration of statewide assessments to all students in selected grades in the content areas of reading, writing, mathematics, and science. The State Board of Education adopted a policy of inclusion of all students who could participate, with appropriate accommodations, in the assessment program. This was consistent with federal legislation (U.S. Department of Education, Individuals with Disabilities Education Act, 1997 Amendments and the Improving America's Schools Act of 1994), which states that all children are expected to meet challenging standards set by their own states. In the first two years of implementation of the CSAP (Spring, 1997 and 1998), comparable assessments in English and Spanish were administered in grade 4 in 1997 and 1998 and grade 3 in 1998.

Purpose of the Study

By the Fall of 1998, it became apparent that support for comparable state assessments in Spanish at the elementary level was waning, and that it was likely that only an English-language version would be available for the grade 5 mathematics assessment in Fall 1999. Drawing on the results of studies indicating that students' language background and the linguistic complexity of an assessment affect student performance on a grade 8 mathematics assessment (Abedi, Lord, & Plummer, 1997; Abedi, Lord, & Hofstetter, 1998; Abedi, Hofstetter, Baker, & Lord, 1998), a joint study was undertaken by the Colorado Department of Education (CDE) and the Center for Research on Student Standards and Testing (CRESST). The purpose of the study was to provide empirical results to inform decisions regarding language accommodations for students who are English language learners (ELL) from all language backgrounds and students with other special needs in the development of Colorado's statewide assessment program.

After discussion of the impact of language on student performance in mathematics and description of the design of this study, its methodology and field operations, analyses, results, and conclusions, this paper describes how the results of the study had a direct impact on the development and construction of the new assessment of fifth-grade mathematics for the Colorado Student Assessment Program.

Review of the Literature

Literature has documented the importance of language in student performance on assessments in content-based areas such as mathematics (see, for example, Abedi, Lord, and Plummer, 1995; Abedi, Lord, and Hofstetter, 1998; Aiken, 1971; Aiken, 1972; Cocking and Chipman, 1988; De Corte, Verschaffel, and DeWin, 1985; Jerman and Rees, 1972; Kintsch and Greeno, 1985; Larsen, Parker, and Trenholme, 1978; Lepik, 1990; Mestre, 1988; Munro, 1979; Noonan, 1990; Orr, 1987; Rothman and Cohen, 1989; Spanos, Rhodes, Dale, and Crandall, 1988). Children perform 10 percent to 30 percent worse on arithmetic word problems than on comparable problems presented in numeric format (Carpenter, Corbitt, Kepner, Linquist, & Reys, 1980). The large gap between the performance of English language learners (ELLs) and native English speakers on math items with high language demand strongly suggests that factors other than mathematical skill contribute to success in solving word problems (Cummins, Kintsch, Reusser, & Weimer, 1988). For example, when the language of instruction is the students' weaker language, bilingual students showed lower performance, but

they score higher on the version of a math in their native language (e.g., Cocking and Chipman, 1988; Macnamara, 1966).

Text comprehension is a crucial step in the problem-solving process. In a review of studies on language and mathematics, Aiken (1971, 1972) showed significant correlations between high reading ability and high arithmetic problem-solving ability. Rothman and Cohen (1989) indicated that there is a link between language and the vocabulary of mathematics. Ginsburg (1981) found that the vocabulary children have for expressing math and number concepts differs widely. Cummins et al. (1988) claim that word problems constitute tests of verbal sophistication as well as logico-mathematical knowledge. In other studies as well, changing the language of the problem to make the relationships clearer raised student performance (De Corte, Verschaffel, & DeWin, 1985; Riley, Greeno, & Heller, 1983). The results suggest that certain problems may be difficult for some children because they cannot interpret key words and phrases in the problem text. In addition, some linguistic factors may present special difficulties for non-native speakers of English. Spanos et al. (1988) identified potential difficulties with comparative structures, prepositional phrases, article usage, conditionals, long nominals (noun phrases), and passive voice constructions, as well as unfamiliar cultural content and vocabulary items that have different meanings in the mathematics context. Thus, some cultural factors may also affect the way language and mathematics interrelate.

Among other factors indicative of potential linguistic complexity, an obvious candidate is the length of the problem statement. Lepik (1990) looked at a large number of structural and linguistic features in algebra word problems, including word length, number of words, number of sentences, and sentence length. He found the highest correlation between the number of words in the problem statement and problem-solving time; however, he did not find a significant relationship between any of the linguistic variables he considered and the proportion of correct responses. None of the variables correlating length of prompt with student achievement reached significance in Lepik's study, in contrast to the findings of Jerman and Rees (1972), who found a significant correlation between length of prompt and number of correct responses.

MacDonald (1993) examined written sentences in which there was a need to resolve lexical and grammatical category ambiguities (the word *trains,* for example, can be a noun or a verb). Her results show that word frequency in the lexicon, both within and across grammatical categories, was one of the primary factors contributing to the resolution of such ambiguities. An alternative to reliance on standardized passages for measuring comprehension is the Cloze procedure, in which words in a passage are deleted at intervals, for example, every fifth word (Taylor, 1953). Using Cloze items to assess comprehension difficulties of reading passages, Bormuth (1966) identified a number of linguistic variables that correlate with passage difficulty, including mean word depth, the ratio of verbs to conjunctions, and letter redundancy, as well as words per sentence and syllables per word. The concept of word depth is a sophisticated measure of syntactic complexity based on a tree diagram of the linguistic structure of a sentence (MacGinitie & Tretiak, 1971; Wang, 1970; Yngve, 1960). Bormuth found a correlation of .86 between sentence length and word depth; consequently, sentence length was supported as an index of complexity in computing readability. Thus, although sentence length may not be a cause of difficulty, it serves as a convenient index for syntactic complexity and can be used to predict comprehension difficulty.

Study Design

The research that was most directly influential on the present study was a CRESST study of NAEP math performance and accommodations and interactions with student language background (Abedi, Hofstetter, Baker, & Lord, 1998). In this study, a sample of grade 8 students in southern California were administered three mathematics test forms comprised of 35 items from the 1996 NAEP Grade 8 mathematics assessment. Two accommodations, extra time and a glossary of non-mathematical terms, were incorporated, resulting in five experimental conditions:

1. original wording of the math items retained and administered with extra time;
2. original wording of the math items retained and administered without extra time;
3. original wording of the math items retained, a glossary of non-mathematical terms provided, and administered with extra time;
4. original wording of the math items retained, a glossary of non-mathematical terms provided, and administered without extra time; and
5. modified wording to simplify non-math vocabulary and reduce complex syntactic structures; administered without extra time.

The results of the Abedi et al. (1998) study indicated that for students in the eighth grade, providing a glossary of non-math terms and allowing extra time increased the scores of ELL and non-ELL students alike. For ELL students, mean scores were highest on the glossary version with extra time allowed, followed by the linguistically simplified version.

Since there was no basis for assuming that these results would hold for elementary school students (in the fourth and fifth grades), a joint research study of the effect of test form on the math performance of elementary school students was conducted in schools throughout Colorado. In addition, the Abedi et al. study did not examine the effect of test form on the math performance of students with disabilities (SD), the majority of whom are assessed by the regular (i.e., not alternate) assessments in the Colorado Student Assessment Program (CSAP). The present study was designed to produce empirical results to inform the development of the grade 5 mathematics assessment, which would be administered to virtually all fifth grade students for the first time in the Fall of 1999[1]. Three test forms were administered: Original English form, Simplified English form, and Original English with Glossary. In this study, extra time was an available accommodation for each of the three forms. These test forms are described in a later section.

Research Hypotheses

The research hypothesis that framed the design of this study is stated below in both its null and alternative forms:

H_0: There is no significant difference in the mathematics performance of students administered mathematics test forms that differ only in linguistic complexity.

H_1: Students who are English language learners will score significantly higher on the Simplified English form and on the Original version with a Glossary than on the form containing the original version of the items.

H_2: Students with disabilities will score significantly higher on the Simplified English form and on the Original version with a Glossary than on the form containing the original version of the items.

[1] Of the 54,875 students enrolled in Colorado public schools in Fall 1999, only 724 (1.3%) were not tested because they could not read English (N = 342) or because their IEPs stated that they were working toward individualized standards (N = 382).

Grade Tested and Source of Items

Released grade 4 mathematics items from the National Assessment of Educational Progress and other assessments were selected by assessment and mathematics education staff of the Colorado Department of Education. Items were selected to meet the test specifications for the Grade 5 Fall CSAP Mathematics Assessment; the items were selected based on alignment with the Colorado Model Content Standards in Mathematics, test format, and range of difficulty.

Language Accommodations and Test Forms

Three test forms were developed using 24 items (16 selected-response and 8 constructed-response, for a total of 34 possible points) that met the specifications of the mathematics assessment framework for the CSAP. Each test form contained the same items, and thus, did not differ in the cognitive demands with respect to mathematics placed on the examinees; however, the test forms differed in the *linguistic* demands placed on the examinees. In addition to the mathematics assessment, each student was administered a measure of English proficiency in reading and teachers completed student language and background questionnaires for all assessed.

The three test forms administered were:

1. the original English version of the items ("Original");
2. a "simplified English" version of the test booklet ("Simplified"); and
3. a "glossary" version, constructed of the original English version of the items and definitions of non-mathematics vocabulary that appeared unnecessary difficult and concepts that might be unfamiliar in other *languages* ("Glossary").

The "Original English" test booklet contained the original wording of the NAEP and other assessment items; no changes were made to structure, format or content of the items. In the "Simplified English" test form, changes were made only to linguistic structures and non-mathematics vocabulary, so that the original mathematics content and mathematics vocabulary were retained. The "Glossary" version provided definitions of non-mathematics vocabulary thought to be unnecessarily difficult for students with disabilities or limited English proficiency. These definitions were presented in shadowed boxes directly on the page where the word occurred. Below is an example of one item in which the linguistic demands varied across the three versions, but the mathematics remained the same:

Original version:

A certain reference file contains approximately six million facts. About how many thousands is that?

Simplified version:

Mack's company sold six million hamburgers. About how many thousands is that?

Glossary version:

A certain reference file contains approximately six million facts. About how many thousands is that?

a certain reference file = a folder for papers
contains = holds
approximately = about

Measure of English Reading Proficiency: Language Assessment Scales

In addition to the mathematics assessment, the level of English proficiency of each student was assessed using CTB/McGraw-Hill's Language Assessment Scales (LAS). The LAS is designed to be an accurate and reliable measure of English-language reading and writing skills. Only the Reading Component, Form 2A, of the Reading and Writing assessment was administered in this study. The Reading Component consists of a total of 45 items in the areas or vocabulary, mechanics and usage, fluency, and reading for information.

Student Language and Background Questionnaire

In addition to the math and reading proficiency assessments, teachers completed a language and background questionnaire for each student that gathered information on demographic and background variables such as:

- race/ethnicity,
- gender,
- disability,
- Title 1 eligibility and type,
- SES, and
- migrant status.

The student background questionnaire also gathered information on language background and concurrent validity, including:

- first language,
- fluency in reading,
- writing and speaking English,
- language spoken at home,
- grades in mathematics and reading,
- language of instruction in mathematics and reading during the 1998-99 school year, and
- accommodations used during the math assessment.

Design

Three test forms were administered randomly to Spring 1999 fourth-grade students in intact classrooms. These students were a subset of the students who would be administered the first assessment of grade 5 mathematics in the Fall of 1999. Randomization was accomplished by using matrix sampling within classroom. Random assignment of language accommodations within classrooms was necessary to minimize class, teacher, and school effects. As described above, each test booklet contained the same mathematics items, differing only in the linguistic (as opposed to cognitive) demands placed on the students. Proficiency in reading in English was assessed in separate LAS test booklets.

Sample Design

Schools and classrooms were selected to provide adequate sample sizes by strata (language background, English proficiency, and classification into special education or the general curriculum) and to provide a reasonable geographic distribution of students within Colorado. Schools and classrooms that had high proportions of ELL students and/or special education students were oversampled. The sample was a 4 x 3 design, with four categories of student classification and three treatments (i.e., test forms), resulting in a 12-cell matrix. The design specified approximately 100 students per cell, or a total sample size of approximately 1200 students.

The sample design is illustrated below:

Student Classification		Test Booklet Form		
		Original English	Simplified English	Original English with Glossary
General Education, non-ELL				
ELL	Spanish			
	Non-Spanish			
Special Education				

n = approximately 100 students/cell

Methods and Procedures

Participants

Data were collected from 1198 fourth-grade students during January 1999. Participating students came from fourteen school districts, 26 schools, and 60 classrooms across Colorado. Within classrooms, all students were sampled. From among the schools in these fourteen districts that met either the high ELL (as indicated by the English Language Proficiency Act, or ELPA, fall count) or high special education criteria (or both), to the extent possible we attempted to select schools that provided a range of socio-economic status (using percentage of students receiving free and reduced-price lunch as a proxy) and ethnic diversity. Table 1 contains the demographic characteristics of the participating schools.

Table 1. Demographic Characteristics of Participating Schools[2]

Descriptive Statistics	% ELPA	% Special Education	% Free or Reduced-price Lunch	% Non-white
N	26	26	26	26
Mean	12.7	13.0	54.8	52.1
Median	11.5	12.5	61.0	51.0
Std. Deviation	9.3	7.0	23.6	24.9
Minimum	1.00	.00	13.00	13.00
Maximum	32.00	29.00	93.00	96.00

Limited English proficient students from both Spanish-speaking and non-Spanish-speaking language backgrounds were over-sampled. The first language of 90 percent of Colorado's ELL population is Spanish. In addition, several schools that contained high proportions of non-Spanish-speaking ELL students were purposefully selected in order to determine the effects of the language accommodations on the performance of students who had no possibility of ever being assessed statewide in their native language in Colorado.

Within districts, district assessment coordinators assisted in recruiting the selected schools. Within schools, we requested that all fourth grade classrooms participate. Between one and four class-

[2] Data from Colorado Department of Education Fall 1997 enrollment records.

BEST COPY AVAILABLE

rooms per school participated. Gaining cooperation from some buildings and districts was difficult because the fourth-graders were scheduled to participate in the statewide testing program approximately five weeks after this study. It was necessary for the research team to administer the tests in three of the districts in order to gain these districts' and schools' participation.

Schools were initially sampled from thirteen districts. As the initial 23 schools and 57 classrooms were recruited and we determined the number of English language learners who would participate in the study, we realized that the number of non-Spanish-speaking limited English proficient students would not be sufficient for the study. (Colorado ELPA records do not identify English language learners' primary language so we could not determine how many non-Spanish-speaking second language learners were included in the sample until we contacted the schools.) As a result, we recruited three additional schools that served a sizable East Asian immigrant population from a fourteenth district. Table 2 contains information about the 1198 participating students' dominant home language and Table 3 contains information about their membership in special education programs. As a result of the over-sampling, 22 percent of our sample of English language learners were non-Spanish-speaking[3]. This is more than double the proportion in the ELL population in Colorado's public schools, where approximately 10 percent speak a language other than Spanish. It is somewhat of a curiosity that our sample contained a slightly *smaller* proportion of special education students than is found in the state as a whole. Over-sampling of schools and classrooms that had high proportions of special education students netted only 8.7 percent of study participants who were reported as enrolled in special education programs by their teachers. Teachers did not provide this information for 17 percent of the study participants. It is possible that if the special education status of these students were known, their proportion in the sample might meet or exceed that of the state public school population (approximately 11.3 percent).

Table 2. Home Language of Study Participants

Language		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	English	772	64.4	77.0	77.0
	Spanish	180	15.0	17.9	94.9
	Other Language	51	4.3	5.1	100.0
	Total	1003	83.7	100.0	
Missing	System	195	16.3		
Total		1198	100.0		

Table 3. Participants' Membership in Special Education Programs

Special Education		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	908	75.8	91.3	91.3
	No	87	7.3	8.7	100.0
	Total	995	83.1	100.0	
Missing	System	203	16.9		
Total		1198	100.0		

[3] Estimates are based on data from the student background questionnaires that were completed by the teachers. Teachers provided this information for 74 percent of the sampled students; data on home language were not provided for 16 percent of the students.

BEST COPY AVAILABLE

Field Operations

Coding and Tracking. The research team contacted each school to obtain exact enrollment figures for each participating classroom. A set of unique student identification (ID) numbers was generated and each classroom was assigned a set of ID numbers based on enrollment plus a ten percent overage for new students.

A color-coded folder was prepared for each ID number. All items in each folder were precoded with this unique ID number. Each folder consisted of the following items:

- one version of the math assessment;
- LAS assessment;
- LAS answer sheet; and
- student background questionnaire response sheet.

Using this system allowed the research team to assign every ID number to a specific version of the math assessment and ensured that each classroom received approximately the same number of each version of the math assessment. This tactic, in conjunction with the instructions to teachers to randomly assign the ID numbers to the students in their class using the last name of the student or some other random assignment method, provided a random spiral of the three versions of the math assessments within each classroom.

Additionally, a packet was prepared for each teacher. Each teacher's packet contained the student folders for their classroom and a teacher key for associating each ID number with a student's name. The teacher key provided the only link between the ID numbers and the names of the students taking those particular tests. Teachers were instructed to keep the teacher keys and not to return them to CDE. Assessment results were returned to teachers by ID number only. In this way, confidentiality was ensured. Results from both the math assessment and LAS were returned to teachers in Spring 1999.

Assessment Training. The research team provided two, ninety-minute assessment administration training sessions and encouraged all of those who would be administering the tests to attend. Training was held in Denver and Southern Colorado. We requested that the math assessment be administered first and that the LAS be administered second, but did not require that these sessions be consecutive or on the same day. We asked that district or building staff administer the assessments so that the classroom teacher would be free to complete the student background and language questionnaire (described below) for all students in the class. Teachers were still responsible for the administration of the LAS reading test. .

Test Administration. The mathematics assessment was a 50-minute test with an additional 10 minutes allowed as part of the standard administration. (Students were allowed to work beyond the 60 minutes if necessary but this was then recorded as an accommodated administration.) The LAS was a 45-minute test. Teachers also completed a language and background questionnaire for each student.

The following accommodations were allowed, provided that the accommodation had been used during instruction for at least three months prior to the assessment:

- reading the math test to the student (this was not allowable for the LAS);
- use of a scribe;
- use of signing or pointing as alternative responses;
- use of an assistive communication device; and
- extended/modified timing or scheduling of administration.

In the participating classrooms, certain students were not included in this study. This included students exempt from testing according to their Individual Education Plan (IEP) and students who were monolingual speakers of another language.

Scoring. Selected-response items were scored in a 0/1 metric (wrong/right). Constructed-response items were hand-scored by elementary teachers, math leaders, and assessment specialists in Colorado using rubrics provided by the original sources of the items and modified or developed by Colorado Department of Education (CDE) staff. These were 2- to 5-point scoring rubrics (0-1 points, 0-2 points and 0-4 points, respectively).

For the constructed-responses (CR) items, scoring guides were obtained from the sources of the original items (i.e., NAEP, TIMSS, MARS, and the New Standards Project). These rubrics were modified by CDE mathematics curriculum and assessment specialists to match the generic rubrics being developed for the constructed-response items in the CSAP Grade 5 Mathematics Assessment. Scoring of student responses was accomplished by a scoring team consisting of 16 selected elementary teachers and math leaders and assessment specialists within Colorado. Three members of this team were bilingual (Spanish) elementary teachers and one was the LEP specialist at CDE. Many of these elementary teachers and math leaders had served on the Elementary Mathematics Team for the development of the Colorado Model Content Standards in Elementary Mathematics and as members of the Colorado team for the New Standards Project.

In preparation for a scoring training session and the actual scoring of the test, the CDE research team selected anchor papers and training "practice papers" for each CR item from actual student work that illustrated the various score points in the rubrics. These pre-scored anchor papers and practice papers were used in the training of the scorers, and provided the basis for calibration. A half-day training session was conducted using these anchor and practice papers for calibration. A calibration criterion of 80 agreement with the pre-scored papers was used in the training. Scorers who did not attain the 80 percent agreement with the scoring of the calibration papers received extra training until the criterion was met.

The scoring session was a one-day event with two tables of eight trained scorers each; there was one table leader at each table. After a "refresher" practice scoring and discussion, the scorers completed the scoring of the entire set of eight CR items for each student. If a scorer had any question or concern about the scoring of an item, he or she first discussed it with their table leader. If the table leader could not resolve the issue, the scoring supervisors (CDE mathematics curriculum specialists) were immediately consulted and the issue resolved.

Data Entry. The LAS and responses to the student background and language questionnaire were machine-scanned and written into an ASCII data file. Each student's responses to the mathematics assessment were data-entered by hand. Responses to selected-response items were initially keyed as 0 (omit), 1, 2, 3 or 4, and responses were later recoded into a 0/1 metric (wrong/right). For constructed-response items, the score assigned by the hand-scoring team was data-entered. Due to time and budget constraints, data were keyed a single time.

Findings

This section presents the findings of the descriptive analyses based on teacher responses on the student language and background questionnaires, overall performance on the three mathematics test forms; overall performance on the LAS; and interactions of student background characteristics, including English reading ability, and test form on mathematics performance.

Descriptive Analysis

This section presents the results of the descriptive analyses of the information provided on the student language and background questionnaires.

Due to the over-sampling of schools and classrooms with high proportions of ELL students, as defined by ELPA counts, our sample of 1198 participating students was distributed as follows: 37 percent Hispanic, 36 percent white non-Hispanic, 6 percent Black, 3 percent Asian/Pacific Islander, 2 percent American Indian, and 16 percent not reported. In contrast, the actual racial/ethnic distribution of fourth graders in Colorado public schools in Spring 1999 was 19 percent Hispanic, 70 percent white non-Hispanic, 6 percent Black, 3 percent Asian/Pacific Islander, 1 percent American Indian, and 1 percent race/ethnicity not reported. As expected, the predominant non-English language spoken in the participants' homes was Spanish. Of students whose predominant home language was reported, 18 percent spoke Spanish, and five percent spoke other languages in their homes. According to the teachers of the sampled students, fewer than 2 percent could not read in English at all, nearly 16 percent read in English "not well," and 37 percent read "fairly well". However, 98 percent of the students received their mathematics instruction entirely in English in the 1998-99 school year, while 96 percent received their reading instruction entirely in English. Teachers also were asked about these students' mathematics and reading grades during the preceding 1998-99 Fall semester. Of those whose math grades were reported, 63 percent received an A or B, 27 percent received a C, and 10 percent received a D or F. In reading, 61 percent received an A or B, 30 percent received a C, and 9 percent received a D or F in the 1998-99 Fall semester. Perhaps because of the sampling criterion of high ELPA counts, 47 percent of the students in this study were in schools that received schoolwide Title I services. Another four percent received targeted Title I services in reading, mathematics, or both. Of the students for whom data were reported, 64 percent received free or reduced-price lunch. Teachers reported that nine percent of these students had a disability identified on their IEP.

Performance on the Forms of the Mathematics Assessment

Results of this study indicated a definite floor effect, regardless of the test form administered. That is, the mathematics assessment was extremely difficult. Mean raw score across all forms was 12.13 out of a possible 34 points. Mean score by test form is shown in Table 4.

Table 4. Mean Mathematics Score by Test Form

Test Form	Mean	Std. Error of Mean	Standard Deviation
Original version	12.14	0.24	4.31
Simplified version	12.13	0.24	4.24
Original with glossary	12.11	0.26	4.42
Total	12.13	0.14	4.32

Several two-factor analysis of variance (ANOVA) procedures were performed to evaluate the impact of linguistic modification of test form and English proficiency on student performance in mathematics. The two-factor analyses of variance (ANOVAs) for the total sample indicated significant main effects for the English proficiency measures -- LAS score, reading grade in class, teacher's perception of how well the student reads in English, and language spoken in the home -- but not for the test form factor. No interaction of test form and English proficiency was found for any of the measures. We believe that this negative result for variation in test performance by test form is due solely to the extreme difficulty of the test, which was comprised primarily of released

BEST COPY AVAILABLE

NAEP items[4]. Item analysis indicated that across all forms, fully half of the items (12 out of 24) had p-values less than 0.33, and six of those had p-values of less than 0.20. Table 5 illustrates the minor amount of variation in p-values among the three test forms.

Table 5. Percentage of Test Items Exhibiting P-Values < 0.33 and < 0.20

Test Form	Percent P < .33	Percent P < .20
Original English	58.3	25.0
Simplified English	50.0	20.8
Original English with Glossary	50.0	29.2
All Test Forms	50.0	25.0

Although the ANOVA indicated no interaction between the two factors, LAS score and test form, t-tests of differences in mean math performance on the three test forms within LAS quintiles shows that students in the second quintile performed significantly better on the Glossary form than they did on the Original form (mean = 11.04 vs. 9.68). Students in the fourth quintile performed significant better on the Simplified version than on either the Original version or the version with the Glossary (mean = 13.76, 12.20, and 12.34, respectively). These results are shown in Table 6.

Table 6. Effect of Test Form and English Proficiency on Math Performance

LAS Quintile	Test Form	Mean	N
1	No difference	9.2	143
2	Glossary	11.0	156
3	No difference	11.6	113
4	Simplified	13.8	193
5	No difference	14.9	227

Similar results were found for students receiving the lowest grades in reading. Students who received D's or F's in reading the semester preceding the assessment period performed significantly better on the Glossary form or the Simplified form than they did on the Original version.

The ANOVA main effects indicated that students with disabilities exhibited lower performance overall than did students without disabilities, but no effect of test form or interaction between the two variables. Another ANOVA, consistent with prior findings on socioeconomic status (SES), showed a main effect for students receiving free or reduced lunch; however no main effect for test form or interaction were found. Students receiving Title I services did not perform significantly differently from students who did not receive such services. This may be due to classification of 92 percent of the "Title I students" as such because they were enrolled in schools in the schoolwide Title I program.

Results on the LAS

Since only the selected-response portions of the LAS Reading Component was used to measure proficiency in reading in English, student results could not be scaled and reported in the standard LAS metric. Student results were reported in terms of raw score (i.e., number correct), and quintile. The mean score on the LAS was 39 (out of a possible 45) with a standard deviation of 6.6. T-tests

[4] Although we attribute the generally poor performance on each of the test forms to the difficulty of the NAEP items, it is important to note that 4th grade students in Colorado usually perform comparatively well on NAEP. In the 1996 assessment, 22 percent of Colorado 4th graders were at or above proficient, compared to 20 percent of students nationally (Shaughnessy, Nelson, & Norris, 1997).

BEST COPY AVAILABLE

for differences in mean math score within LAS quintile groups revealed no difference in performance on the three math test forms for the least and most English proficient students (i.e., in LAS quintiles 1 and 5). Thus, if a student is very limited in their English proficiency (LAS quintile 1), modifying the linguistic complexity of a mathematics assessment does not help at all; and indeed, this group reported the lowest math scores overall. Similarly, the linguistic complexity of an assessment makes no difference in the mathematics performance of fluent English readers (LAS quintile 5). However, students in LAS quintile 2 scored significantly higher (at $\alpha = .05$) on the form containing the glossary (mean = 11.04) than on either the form containing the original English items (mean = 9.68) or the simplified items (mean = 10.19). Interestingly, math performance also differed by test form for students in LAS quintile 4, with the simplified English form producing the best results (mean of 13.67 vs. 12.20 on the original form and 12.34 on the glossary form. Since ELL students disproportionately receive Title I services, it was postulated that performance on the LAS would vary by receipt of Title I services. However, analysis demonstrated no difference between the two groups. Again, this may be due to the fact that virtually all of the students categorized as Title I are in schools that are in the schoolwide Title I program (92 percent).

Mathematics Performance in the Highest Performing Schools

Because of the obvious floor effect described above, the authors decided narrow the analysis to students in the three top-performing schools. The purpose of this was to lessen the effects of test difficulty, thereby increasing the variance in the score distribution. This procedure was used because of the large positive skew in the distribution and since the focus of the analyses was not on absolute performance. Rather, the focus was on the relative performance among groups of students and the treatment effect.

The total number of students in the three-school sample was 109. This subsample included native-English speakers, native-Spanish speakers, and native-speakers of many other languages. The mean raw scores for the three forms were: 14.16 for the Original English form, 14.03 for the Simplified English form, and 16.26 for the form with the original English and the glossary. The results of the two-way ANOVA using the test forms and LAS quintiles as factors indicated that performance among these three forms was significantly different at the 0.01 level. Except for the most limited English proficient students, students performed best on the test form containing the Glossary. Students with the highest English proficiency performed equally well on the Glossary form and the Simplified English form of the test. The test forms on which students in the five LAS categories demonstrated the best performance are shown in the table below:

Table 7. Effect of Test Form and English Proficiency on Math Performance by Students in the Highest Performing Schools

LAS Quintile	Test Form	Mean	N
1	No difference	10.8	13
2	Glossary	16.2	14
3	Glossary	17.0	13
4	Glossary	18.8	27
5	Simplified Glossary	18.1 17.6	42
Total	Glossary	16.7	109

Eleven percent of the subsample was students with disabilities; they performed equally well on the Original and the Glossary forms (means = 13.0), in comparison to the Simplified English form (mean = 10.7). These results are shown in Table 8.

BEST COPY AVAILABLE

Table 8. Effect of Test Form on Mean Math Performance by Disability Status in the Highest Performing Schools

Test Form	Students with Disabilities	Students without Disabilities	All Students
Original	13.0	14.6	14.3
Simplified	10.7	14.5	14.2
Glossary	13.0	17.1	16.7
Total	12.4	15.3	15.0

Table 8 also indicates that the math performance of non-disabled students in the highest performing schools was significantly higher on the Glossary form than on the Original or Simplified English forms. In these top-performing schools, there were too few students receiving Title I services or free or reduced lunch to support comparisons of test performance among these students.

Conclusions

The initial conclusion that can be drawn from these data is that students' performance on mathematics assessments with high proportions of word problems is directly related to their proficiency in reading in English. Examination of math performance within LAS category suggests that simplification of linguistic structures and the addition of a glossary for non-mathematics vocabulary to a math assessment results in better performance by English language learners and other students who are not good readers.

Although our hypotheses that students who are English language learners or who have disabilities would score significantly higher on the modified test forms was not supported by the overall analyses, subsequent analysis of the results from students in higher performing schools indicate that these hypotheses are tenable. When the effect of test difficulty was controlled, it was found that all but the most limited English proficient students, including students with disabilities, performed best on the Glossary form of the test. Students with the highest English proficiency performed equally well on the Glossary form and on the Simplified English form. Thus, all groups did at least as well, if not better, on the Glossary form of the assessment instrument as on the Simplified form.

These data suggest that linguistic simplification or clarification of the vocabulary of mathematics word problems can benefit *virtually all students*. Thus, we feel that unnecessarily complex linguistic structures or difficult vocabulary in a mathematics assessment introduces non-construct-related variance that can be removed by careful attention to construction of the assessment to measure the construct of math knowledge – not reading ability.

A secondary conclusion that can be drawn from these data is that the released NAEP and other items selected for this study were too difficult for fourth-grade students. Although we requested released grade 4 items from the NAEP assessment, subsequent examination of released blocks of NAEP items revealed that several items used this study had been administered to *grade 8* students in the 1996 assessment. Not surprisingly, these items had the lowest p-values. Thus, we attribute the overall lack of interaction of the linguistic complexity of test form and English proficiency to the extreme difficulty of the test and the resulting lack of variation in the score distribution.

Implications for the Colorado Student Assessment Program Mathematics Assessment

Based on the results of this study (and on common sense), several decisions were made in constructing the grade 5 mathematics assessment of the Colorado Student Assessment Program.

First, every attempt was made to avoid unnecessary linguistic complexity. All of the potential test items in the item pool supplied by the state's assessment contractor were reviewed for linguistic features that appear to contribute to text difficulty but were not related to the math content of the item. Most of the items in the assessment item pool were subsequently modified to meet this criterion. In addition, definitions of non-mathematical words were provided, where appropriate, underneath the test item. In no case was the *mathematical* complexity compromised.

References

Abedi, J., Lord, C. & Hofstetter, C. (1998a). *Impact of Selected Background Variables on Students' NAEP Math Performance*. Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (1998b). *Performance and Test Accommodations; Interactions with Student Language Background*. Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C. & Plummer, J. (1995). *Language Background as a Variable in NAEP Performance: NAEP TRP Task 3D: Language Background Study*. Los Angeles: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Aiken, L. R. (1971). Verbal factors and mathematics learning: A review of research. *Journal for Research in Mathematics Education, 2*, 304-13.

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Education Research*, 42(3), 359-85.

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79-132.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Jr., Linquist, M. M., & Reys, R. E. (1980, September). Solving verbal problems: Results and implications from national assessment. *Arithmetic Teacher, 28*, 8-12.

Cocking, R. R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and Cultural Influences on Learning Mathematics*, pp. 17-46. Hillsdale, NJ: Erlbaum Associates.

Colorado Department of Education (1998). *Colorado School Laws 1998*. Denver, CO: Colorado Department of Education.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77*(4), 460-470.

Ginsburg, H. (1981). The clinical interview in psychological research on mathematical thinking: Aims, rationales, techniques. *For the Learning of Mathematics, 1*(3), 4-11.

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. *Educational Studies in Mathematics, 4*, 306-323.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*(1), 109-129.

Larsen, S. C., Parker, R. M., & Trenholme, B. (1978). The effects of syntactic complexity upon arithmetic performance. *Educational Studies in Mathematics, 21*, 83-90.

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21*, 83-90.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32*, 692-715.

MacGinitie, W. H., & Tretiak, R. (1971). Sentence depth measures as predictors of reading difficulty. *Reading Research Quarterly, 6*, 364-377.

Macnamara, J. (1966). *Bilingualism in Primary Education*. Edinburgh: Edinburgh University Press.

Mestre, J.P. (Fall, 1984). The problem with problems: Hispanic students and mathematics. *Bilingual Journal*, 15-32.

Munro, J. (1979). Language abilities and math performance. *Reading Teacher, 32*(8), 900-915

Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care, 54*, 57-81.

Orr, E. W. (1987). Twice as Less: Black English and the Performance of Black Students in Mathematics and Science. New York: W. W. Norton.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.). The *Development of Mathematical Thinking* (pp. 153-196). New York: Academic Press.

Rothman, R. W., & Cohen, J. (1989). The language of math needs to be taught. *Academic Therapy, 25*(2), 133-42.

Shaughnessy, C. A., Nelson, J. E., & Norris, N. A. (1997). *NAEP 1996 Mathematics Cross-state Data Compendium for the Grade 4 and Grade 8 Assessment*. Washington, DC: National Center for Education Statistics.

Spanos, G., Rhodes, N. C., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Eds.), *Linguistic and Cultural Influences on Learning Mathematics* (pp. 221-240). Hillsdale, NJ: Erlbaum Associates.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415-533.

Wang, M. D. (1970). The role of syntactic complexity as a determiner of comprehensibility. *Journal of Verbal Learning and Verbal Behavior, 9*, 398-404.

Yngve, V. H. (1960). A model and hypothesis for language structure. *Proceedings of the American Philosophical Association, 404*, 444-466.